S-5

Mining the Traffic Conditions via Twitter based on Rough Set Theory

Septianusa, Maulina Supriyaningsih, Ayu Septiani, RB Fajriya Hakim

Department of Statistics, Faculty of Mathematics and Sciences, Universitas Islam Indonesia
hakimf@fmipa.uii.ac.id

Abstract:

This paper show the traffic conditions from several lane towards or out of the port of Tanjung Priok. As the biggest port for exports and imports to support the economy of Indonesia, then jam to or out of the port would be very detrimental. By using data retrieved from twitter, we give an overview of traffic conditions using if-then rules of rough set theory which is expected to give people or local government that are interested in these area a valuable information. Some interesting rules are obtained with their certainty and coverage factors.

1. Introduction

Although Indonesia is a developing country but this country is one of the highest level of twitter usage [1]. The high usage of twitter could be used to gain new information and knowledge quickly. One of the advantage use of twitter is it can be used to broadcast about information regarding a particular road congestion. The road which pays an attention to the authors is the road towards or out of the port of Tanjung Priok. As we know, Tanjung Priok is the largest seaport in Indonesia. As the biggest port for exports and imports to support the economy of Indonesia, then jam to or out of the port would be very detrimental. The impacts of congestion at Tanjung Priok areas are not only detrimental to the economy, but also disrupt the daily lives of other road users. With the calculation of operating cost 1 million rupiahs per day per truck per trip, then of 18,000 trucks operating in those region reached 18 billion rupiahs per day, assuming the roads are not jammed [2]. However, if half of the truck was not in operation because of bad roads, the losses reached 9 billion rupiahs per day.

Seeing the huge losses incurred due to jams in those region, the authors would like to give an overview of the jam on those areas, to help communities and governments to make decisions on an individual scale and policies on a wider scale. Overview which is used to indicate jam here using if-then rules from the rough set theory. The theory of rough sets introduced by Pawlak [3] and has been used in many fields related to uncertain data. The data used in this study is the data of the users twitter tweets well as individuals, private institutions and government agencies that provide information on road conditions around the harbor of Jakarta.

2. Related Research

The research done by Kosala, et al.[4], attempted to extract road traffic information from Twitter timelines for real time mapping. They proposed an algorithm to measure the traffic information confidence level for the real time traffic mapping system. Their system could be used to serve its intended purpose [4].

The paper done by Endarnoto, et al., used the information extraction technique to get the data of traffic from the Twitter account of the TMC Polda Metro Jaya, so that the traffic information can be presented in map view as a mobile application of Android. Early experiment with limited vocabulary and rules has showed promising result [5].

The research done by He, et al., examined whether it is possible to use the rich information in online social media to improve longer-term traffic prediction. They first analyze the correlation between traffic volume and tweet counts with various granularities. Then they propose an optimization framework to extract traffic indicators based on tweet semantics using a transformation matrix, and incorporate them into traffic prediction via linear regression. Experimental results using traffic and Twitter data originated from the San Francisco Bay area of California demonstrate the effectiveness of their proposed framework [6].

3. Research Methodology

Tweets data were obtained from several accounts that observe the traffic conditions in Jakarta. Those accounts are @infoll, @infolantas, @infomacetcom, @TMCPoldaMetro, @PTJasaMarga and @lewatmana, the accounts selected as active in preaching the traffic conditions in Jakarta. Retrieval and processing research data using statistical software package R and the package "twitteR" [7]. Data collection was conducted twice, at the first stage is obtained as 3601 tweets on 3 December – 18 December 2013. While in the second phase, obtained 2887 tweets on 19 December – 29 December 2013.

Tweets that has been obtained is then transformed into a table that is broken down by day, time, and the status of the traffic lane. The table is then analyzed using the methods of rough sets to get the rules regarding traffic conditions in the area of Tanjung Priok. Variable used in this researchs are:

a. Variable Day
   Tweets grouped by day with the date there, namely: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday.
b. Variable Time
   Tweets grouped by time (western Indonesian time (GMT), +7 from server logs Twitter.com time) at the time of road condition information provided by the respective twitter accounts.
   Grouping time is Midnight – Morning (T1) (00:00 to 06:00), Morning – Afternoon (T2) (06:00 to 12:00 pm), Afternoon – Evening (T3) (12:00 to 18:00), Evening – Midnight (T4) (18.00-24.00)
c. Variable Lane
   Variable lane is a lane that is commonly used by four-wheeled vehicles or more. The intended lane is (source for all pictures are from Google Map):

Pluit – Tanjung Priok Lane (Figure 3.1)

Figure 3.1 Lane of Pluit-Tanjung Priok

- A (Pluit area) to B (Ancol area) is the lane Prof Sedyatmo Toll – Port Toll
- B (Ancol area) to C (Tanjung Priok) is the lane Poll Toll – R.E. Martadinata Street

Cawang - Tanjung Priok Lane (Figure 3.2.)

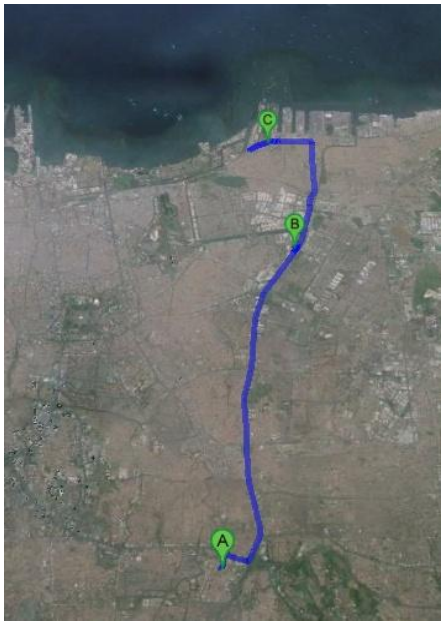

Figure 3.2 Lane of Cawang Area-Tanjung Priok

- A (Cawang Area) to B (Sunter Timur Area) is lane of Ir. Wiyoto Wiyono Toll
- B (kawasan Sunter Timur) to C (Tanjung Priok) is Enggono and Yos Sudarso Street

Cakung – Tanjung Priok Lane (Figure 3.3)

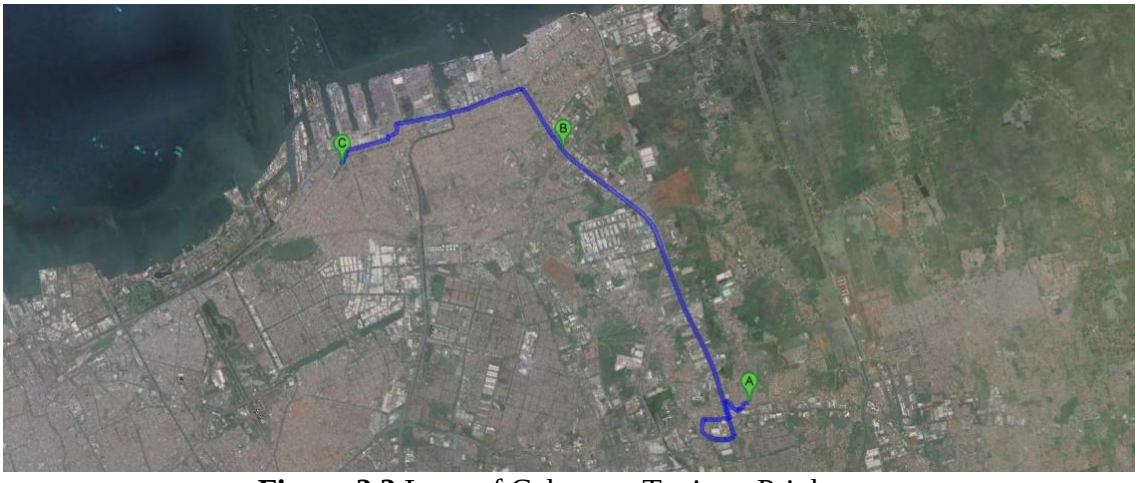

Figure 3.3 Lane of Cakung – Tanjung Priok

- A (Cakung Area) to B (Cakung Cilincing) is lane of Outer Ring Road Toll and Cakung Cilincing Raya Street
- B (Cakung Cilincing) to C (Tanjung Priok) is lane of Kalibaru Barat and Enggono Street

d. Variable Traffic Condition

Traffic conditions are quantifying by different congestion levels, "Macet" (heavy traffic jam), "Tersendat" (light traffic jam), "Padat merayap" (heavy traffic), "Ramai Lancar" (light traffic), and "Lancar" (very light traffic).

Based on observational data obtained as 6488 tweets, all using Indonesian and since we cannot give assurance that the tweets using proper grammar Indonesian, then some natural language processing (NLP) techniques might be rendered useless to tackle the problem. This raw data are stored using extension .csv and use the "search" menu in the spreadsheet aplication software to find the tweets that contain keywords which is names of lane that towards or out of the Port. From this raw data after some complicated prepocessing methods we get 1058 relevan tweets.

4. Result and Discussion

The purpose of this study is to explore a useful rule of piles of tweets about traffic conditions in the area of Tanjung Priok using Rough Set Theory. Before any data is analyzed using the method of Rough Set, there are preprocessing stages. The aims of preprocessing stage is to handle the missing data and to classify them according to their

category. Two factors measurement of Rough Set Theory which are certainty factor and coverage factors [3] will be used,

1. *Certainty Factors*

$$\pi(\psi|\phi) = \frac{\text{number of all cases satisfying } \phi \text{ and } \psi}{\text{number of all cases satisfying } \phi}$$

2. *Coverage Factors*

$$\pi(\phi|\psi) = \frac{\text{number of all cases satisfying } \phi \text{ and } \psi}{\text{number of all cases satisfying } \psi}$$

before showing the final result that can be obtained by setting the total tweets about traffic conditions in Tanjung Priok using Rough Set Theory, we will show as an example, the rules which is acquired on Monday traffic for each lane.

Data Reduction based on line A to C (Monday)

The Western line, from A1 (Pluit) to C (Tanjung Priok) through:

- Prof. Sedyatmo Toll – Port Toll
- Port Toll – R.E. Martadinata Street

Table 3.1 Data Reduction based on line A1 to C

Day	Time	Pluit - Tanjung Priok	N	Certainty	Coverage
Monday	T1	Ramai Lancar	3	0,75	0,125
Monday	T1	Padat Merayap	1	0,25	0,047619
Monday	T4	Ramai Lancar	7	0,875	0,291667
Monday	T4	Padat Merayap	1	0,125	0,047619
Monday	T3	Tersendat	2	0,090909	0,333333
Monday	T3	Ramai Lancar	7	0,318182	0,291667
Monday	T3	Padat Merayap	12	0,545455	0,571429
Monday	T3	Macet	1	0,045455	0,333333
Monday	T2	Tersendat	4	0,2	0,666667
Monday	T2	Ramai Lancar	7	0,35	0,291667
Monday	T2	Padat Merayap	7	0,35	0,333333
Monday	T2	Macet	2	0,1	0,666667

Table 3.1. contained six columns which are Day, Time, Road Condition (From Pluit to Tanjung Priok), N (The amount of information provided about the condition of the road), Certainty and Coverage. There are several interesting information from table 3.1., for example, in the first row, said that "the lane conditions from Pluit to Tanjung Priok are Ramai Lancar on Monday Midnight to Monday Morning with certainty 75% and almost 12.5% tweets coverage this conditions.

Some other conditions that are interesting to observe is time (T2) namely Morning to Afternoon, four conditions happened at those time with certainty 2% "Tersendat", 35% "Ramai Lancar", 35% "Padat Merayap" and 1% "Macet" with coverage 66.67%, 29.2%, 33.33% and 66.67% of tweets, respectively. We also could use if-then rules for this information,

If lane Pluit – Tanjung Priok on Monday Morning to Afternoon *then* the road condition could be bad (from "Ramai Lancar" to "Macet").

The South line, from the A2 (Cawang) to C (Tanjung Priok) through:

- Ir. Wiyoto Wiyono Toll
- Enggono and Yos Sudarso Street

Table 3.2. Data Reduction based on line A2 to C (Monday)

Day	Time	Cawang-Tanjung Priok	N	Certainty	Coverage
Monday	T1	Padat Merayap	1	0,25	0,05
Monday	T1	Ramai Lancar	2	0,5	0,083333
Monday	T1	Tersendat	1	0,25	0,076923
Monday	T4	Padat Merayap	7	0,5	0,35
Monday	T4	Ramai Lancar	7	0,5	0,291667
Monday	T3	Padat Merayap	2	0,1	0,1
Monday	T3	Padat Merayap	3	0,15	0,15
Monday	T3	Ramai Lancar	9	0,45	0,375
Monday	T3	Tersendat	6	0,3	0,461538

Monday	T2	Macet	4	0,173913	1
Monday	T2	Padat Merayap	7	0,304348	0,35
Monday	T2	Ramai Lancar	6	0,26087	0,25
Monday	T2	Tersendat	6	0,26087	0,461538

From the table 3.2 we get some information interesting, on Monday Morning to Afternoon (T2) we get 17% certainty "Macet" with coverage 100% tweets state this condition. The rules could be seen as,

If lane Cawang – Tanjung Priok on Monday Morning to Afternoon *then* the road condition could be bad (from "Ramai Lancar" to "Macet")

The Southern line, from the A3 (Cakung) to C (Tanjung Priok) through :

- Lane Outer Ring Road Toll of Jakarta and Cakung Cilincing Raya Street
- Lane Kalibaru Barat and Enggono Street

Table 3.3 Data Reduction based on line Cakung to Tanjung Priok (Monday)

Day	Time	Cakung-Tanjung Priok	N	Certainty	Coverge
Monday	T2	Padat Merayap	4	0,363636	0,5
Monday	T2	Ramai Lancar	7	0,636364	0,368421
Monday	T3	Ramai Lancar	6	1	0,315789
Monday	T4	Padat Merayap	4	0,4	0,5
Monday	T4	Ramai Lancar	6	0,6	0,315789

From table 3.3. the lane condition from Cakung to Tanjung Priok on Monday Afternoon to Evening are 100% certainty "Ramai Lancar" with coverage about 31% of tweets state this condition. Almost all of the tweets in this line provided information about conducive lines.

Data Reduction based on lane A to B, B to C (Monday)
Western Sub Line, from A1 (Pluit) to B1 (Ancol) through:

- Prof. Sedyatmo Toll - Port Toll

Table 3.4. Data Reduction Based on Line Pluit to Ancol

Day	Time	Pluit (Prof Sedyatmo Toll – Port Toll)- Ancol	N	Certainty	Coverage
Monday	T2	Tersendat	5	0,416667	1
Monday	T2	Ramai Lancar	7	0,583333	0,583333
Monday	T3	Ramai Lancar	5	0,333333	0,416667
Monday	T3	Padat Merayap	10	0,666667	0,909091
Monday	T4	Padat Merayap	1	0,333333	0,090909
Monday	T4	Macet	2	0,666667	0,666667
Monday	T1	Macet	1	1	0,333333

On Monday this sub line would be "Macet" on Evening to Midnight (about 66,67% certainty) and 100% certainty "Macet" on Midnight to Morning.

Western Sub Line, from B1(Ancol) to C (Tanjung Priok) through:

- Port Toll - R.E Martadinata Street

Table 3.5. Data Reduction Based on Line Ancol – Tanjung Priok

Day	Time	Ancol-Tanjung Priok	N	Certainty	Coverage
Monday	T2	Tersendat	1	0,0625	1
Monday	T2	Ramai Lancar	15	0,9375	0,9375
Monday	T3	Ramai Lancar	1	0,090909	0,0625
Monday	T3	Padat Merayap	10	0,909091	1

This sub line on Monday Morning to Evening are dominated by "Ramai Lancar" and "Padat Merayap" conditions.

Southern Sub Line, from A2 (Cawang) to B2 (Sunter Timur) through:

- Ir. Wiyoto Wiyono Toll

Table 3.6. Data Reduction Based on Line Cawang – Ir. Wiyoto Wiyono Toll

Day	Time	Cawang-Ir. Wiyoto Wiyono Toll	N	Certainty	Coverage
Monday	T2	Tersendat	6	0,5	0,461538
Monday	T2	Ramai Lancar	1	0,083333	0,2
Monday	T2	Padat Merayap	1	0,083333	0,166667
Monday	T2	Macet	4	0,333333	1
Monday	T3	Tersendat	6	0,5	0,461538
Monday	T3	Ramai Lancar	4	0,333333	0,8
Monday	T3	Padat Merayap	2	0,166667	0,333333
Monday	T4	Ramai Lancar	5	0,714286	0,714286
Monday	T4	Padat Merayap	2	0,285714	0,333333
Monday	T1	Tersendat	1	0,25	0,076923
Monday	T1	Ramai Lancar	2	0,5	0,285714
Monday	T1	Padat Merayap	1	0,25	0,166667

In this sub line, at time (T2) Morning to Evening the lane condition tend to bad condition, 100% coverage that this lane is "Macet".

Southern Sub Line, from B2 (Sunter Timur) to C (Tanjung Priok) through:

- Enggono and Yos Sudarso Street

Table 3.7. Data Reduction Based on Ir. Wiyoto Wiyono Toll – Yos Sudarso Street

Day	Time	Ir. Wiyoto Toll-(Enggono-Yos Sudarso Street) Tanjung Priok	N	Certainty	Coverage

Monday	T2	Ramai Lancar	12	0,666667	1
Monday	T2	Padat Merayap	6	0,333333	0,428571
Monday	T3	Padat Merayap	8	1	0,571429

This sub line on Monday Morning – Evening the lane are "Ramai Lancar" and "Padat Merayap" which is good condition.

Eastern Lane, from A3 (Cakung Area) to B3 (Cakung Cilincing) through:

- Outter Ring Road Toll and Cakung Cilincing Raya Street

Table 3.8. Data Reduction Based on Line Cakung area– Cakung Cilincing

Day	Time	Cakung-Cakung Cilincing	N	Certainty	Coverage
Monday	T2	ramai lancar	3	0,75	0,333333
Monday	T2	Padat Merayap	1	0,25	0,333333
Monday	T3	ramai lancar	3	1	0,333333
Monday	T4	ramai lancar	3	0,6	0,333333
Monday	T4	Padat Merayap	2	0,4	0,666667

This sub line from Morning untill Midnight on Monday are good condition.

Eastern Sub Lane, from B3 (Cakung Cilincing) to C (Tanjung Priok) through:

- Kalibaru Barat and Enggono Street

Table 3.9. Data Reduction Based on Line Cakung Cilincing – Tanjung Priok

Day	Time	Cakung Cilincing-Tj.Priok	N	Certainty	Coverage
Monday	T2	Ramai lancar	4	0,571429	0,4
Monday	T2	Padat Merayap	3	0,428571	0,6
Monday	T3	Ramai lancar	3	1	0,3
Monday	T4	Ramai lancar	3	0,6	0,3
Monday	T4	Padat Merayap	2	0,4	0,4

From Monday Morning to Midnight, this sub line are conducive. All above data are sample on Monday traffic from day to day condition in the Tanjung Priok area.

Data Reduction Based on Rough Set Theory at Lane A to C in All Days

Overall traffics in a weeks based on tweets data with the highest certainty and coverage of each lane (Pluit (A1) – Tanjung Priok (C), Cawang (A2) – Tanjung Priok, Cakung (A3) – Tanjung Priok) are shown in the table 3.10.

Table 3.10. Highest Certainty of Road Condition in All Days of Three Lanes

Lane	Day	Time	Condition	N	Certainty
Pluit-Tanjung Priok	Monday	T4	Ramai Lancar	7	0,875
	Sunday	T2	Ramai Lancar	11	0,785714
	Thursday	T1	Ramai Lancar	6	0,75
	Monday	T1	Ramai Lancar	3	0,75
Cawang-Tanjung Priok	Sunday	T1	Ramai Lancar	1	1
	Tuesday	T4	Ramai Lancar	9	0,75
	Sunday	T4	Ramai Lancar	10	0,714286
	Sunday	T2	Ramai Lancar	9	0,692308
	Friday	T4	Padat Merayap	10	0,666667
	Friday	T1	Ramai Lancar	2	0,666667
Cakung-Tanjung Priok	Sunday	T2	Ramai Lancar	8	1
	Monday	T3	Ramai Lancar	6	1
	Friday	T1	Ramai Lancar	2	1
	Sunday	T1	Ramai Lancar	2	1
	Wednesday	T1	Ramai Lancar	2	1
	Saturday	T1	Ramai Lancar	1	1
	Tuesday	T1	Ramai Lancar	1	1

From the table 3.10 shows that from all data obtained, lane Cakung – Tanjung Priok has higest certainty and state the "Ramai Lancar" condition in almost one week from Midnight to Evening. While lane Pluit – Tanjung Priok on Sunday T2, Monday T1 and T4 and Thursday T1 are "Ramai Lancar". Lane Cawang – Tanjung Priok on Sunday T1 T2 and T4, Tuesday T4 and Friday T1 are "Ramai Lancar". Using decision rules of rough set theory, the rules could be read as,

> *If* lane Pluit – Tanjung Priok on Monday Midnight to Morning and Evening to Midnight *Then* Ramai Lancar with certainty 75% - 87.5%.
>
> *If* lane Pluit – Tanjung Priok on Sunday Morning – Afternoon *then* Ramai Lancar with certainty 78.6%
>
> *If* lane Pluit – Tanjung Priok on Thursday Midnight – Morning *then* Ramai Lancar with certainty 75%

Or we also could say, with certainty more than 75%, lane Pluit – Tanjung Priok will "Ramai Lancar" on Monday Midnight – Morning, Monday Evening – Midnight, Sunday Morning – Afternoon and Thursday Midnight – Morning.

The same goes for other data which could be specified in the if-then rules. Table below (Table 3.11) shows the highest coverage for all three lane to Tanjung Priok

Table 3.11. Highest Coverge for Three Lanes

Jalur	Hari	Waktu	Kondisi	N	Certainty	Coverge
Pluit-Tanjung Priok	Friday	T2	Tersendat	7	0,411765	0,089744
	Tuesday	T2	Tersendat	7	0,318182	0,089744
	Thursday	T2	Tersendat	6	0,315789	0,076923
	Sunday	T4	Ramai Lancar	12	0,571429	0,067416
	Sunday	T2	Ramai Lancar	11	0,785714	0,061798
	Saturday	T4	Ramai Lancar	10	0,555556	0,05618
	Monday	T3	Padat Merayap	12	0,545455	0,084507
	Saturday	T3	Padat Merayap	10	0,416667	0,070423
	Friday	T3	Padat Merayap	10	0,4	0,070423
	Friday	T2	Macet	2	0,117647	0,105263
	Monday	T2	Macet	2	0,1	0,105263
	Wednesday	T2	Macet	2	0,086957	0,105263
Cawang-Tanjung Priok	Friday	T3	Tersendat	6	0,333333	0,101695
	Monday	T3	Tersendat	6	0,3	0,101695
	Monday	T2	Tersendat	6	0,26087	0,101695
	Sunday	T4	Ramai Lancar	10	0,714286	0,064516
	Wednesday	T2	Ramai Lancar	10	0,416667	0,064516
	Tuesday	T4	Ramai Lancar	9	0,75	0,058065
	Tuesday	T2	Padat Merayap	13	0,481481	0,079268
	Friday	T4	Padat Merayap	10	0,666667	0,060976
	Thursday	T3	Padat Merayap	10	0,588235	0,060976
	Monday	T2	Macet	4	0,173913	0,571429

	Thursday	T2	Macet	2	0,1	0,285714
Cakung-Tanjung Priok	Thursday	T2	Tersendat	3	0,230769	0,1875
	Saturday	T3	Tersendat	3	0,214286	0,1875
	Wednesday	T4	ramai lancar	9	0,6	0,059211
	Tuesday	T3	ramai lancar	9	0,6	0,059211
	Wednesday	T2	ramai lancar	9	0,5	0,059211
	Tuesday	T2	ramai lancar	9	0,5	0,059211
	Wednesday	T2	Padat Merayap	7	0,388889	0,097222
	Tuesday	T2	Padat Merayap	7	0,388889	0,097222
	Saturday	T4	Padat Merayap	6	0,545455	0,083333
	Wednesday	T4	Padat Merayap	6	0,4	0,083333

A table 3.11. above is certainty and coverage factor with highest value for each lane toward Tanjung Priok for a week. Some interesting rules could be found from this table.

If lane Pluit – Tanjung Priok on Morning to Afternoon in office day
then road condition will be bad ("Tersendat" and "Macet").

If lane Cawang – Tanjung Priok on Monday Morning to Evening and Friday Afternoon to Evening *then* road condition "Tersendat".

If lane Cakung – Tanjung Priok on Monday and Thursday Morning to Afternoon
then road condition "Macet".

5. Conclusion

This paper is an application of rough set method to mine the traffic data. Rough Set is used to model the rules of traffic conditions in Tanjung Priok, North Jakarta. To obtain these rules, the authors used data from twitter (tweets). Tweets are considered as a representation of data to describe the actual traffic condition. Tweets data are transformed into the frequency table and grouped by day, time, lane and the traffic status. Tweets data then analyzed using the method of rough set to find hidden rules from a pile of data. Then reduction are obtained based on lanes to saw the rules with certainty and coverage of each lane toward Tanjung Priok. The result is expected to provide description and insight of traffic of Tanjung Priok area based on day and

time. It is reasonable for Indonesia as the country with the world's third largest twitter users to enable dissemination and delivery of information more effectively for people who have an interest in the traffic at the region and also for local government in taking policies.

References

[1] HarianTI, Indonesia Jadi Negara Dengan Penetrasi Twitter Tertinggi Di Dunia Saat Ini http://harianti.com/indonesia-jadi-negara-dengan-penetrasi-twitter-tertinggi-di-dunia-saat-ini/ accessed 30 September 2013

[2] Poskotanews, Hindari Priok Macet Akan Mendera http://m.poskotanews.com/2013/09/11/hindari-priok-macet-akan-mendera/ 11 September 2013

[3] Pawlak Z, A Primer on Rough Sets: A New Approach to Drawing Conclusions from Data. *Cardozo Law Review*, Vol.22, No.5-6, 2001, pp.1407-1415.

[4] Kosala, R., Adi, E., Steven, 2012, Harvesting Real Time Traffic Information from Twitter *Procedia Engineering*, *Volume 50*, *2012*, *Pages 1-11*

[5] Endarnoto, S.K., Pradipta, S., Nugroho, A.S., Purnama, J., 2011, Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application, *International Conference on Electrical Engineering and Informatics*

[6] He, J., Shen, W., Divakaruni, P., Wynter, L., and Lawrence, R., Improving Traffic Prediction with Tweet Semantics, *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*

[7] Jeff Gentry, R package: twitteR, 2013